

Grounding as a collaborative process

Benotti, Luciana; Blackburn, Patrick Rowan

Published in:

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics

Publication date:

2021

Document Version

Publisher's PDF, also known as Version of record

Citation for published version (APA):

Benotti, L., & Blackburn, P. R. (2021). Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 515-531). Association for Computational Linguistics. <https://www.aclweb.org/anthology/2021.eacl-main.41.pdf>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy

If you believe that this document breaches copyright please contact rucforsk@kb.dk providing details, and we will remove access to the work immediately and investigate your claim.

Grounding as a Collaborative Process

Luciana Benotti

Universidad Nacional de Córdoba
CONICET, Argentina
luciana.benotti@unc.edu.ar

Patrick Blackburn

Philosophy and Science Studies, IKH
Roskilde University, Denmark
patrickb@ruc.dk

Abstract

Collaborative grounding is a fundamental aspect of human-human dialog which allows people to negotiate meaning. In this paper we argue that it is missing from current deep learning approaches to dialog and interactive systems. Our central point is that making mistakes and being able to recover from them collaboratively is a key ingredient in grounding meaning. We illustrate the pitfalls of being unable to ground collaboratively, discuss what can be learned from the language acquisition and dialog systems literature, and reflect on how to move forward.

1 Introduction

Collaborative grounding is shaped by *constraints* that are not explicit in successful dialog turns. These constraints combine information from world *causal relations*, the *task* under discussion, the *communicative intents* of the dialog partners, and much else besides. They are used to *negotiate meaning*, to decide which beliefs to add to the shared *common ground*, which is constructed using the *joint attention* of the dialog partners to things either real or imagined. Once beliefs are grounded, they cannot be magically ungrounded without further negotiation, for the dialog partners are *committed* to them. But it's tricky:

Human-analogous natural language understanding (NLU) is a grand challenge of artificial intelligence, which involves mastery of the structure and use of language and the ability to ground it in the world. (Bender and Koller, 2020)

What does “the ability to ground it in the world” involve? Implicit constraints that shape grounding become explicit when communication starts to break. We claim that *making mistakes* and being able to recover from them *collaboratively* is a key ingredient of the ability to ground.

We proceed as follows. We first discuss obstacles to current research on dialog and interactive systems (Section 2) and then present working definitions of *collaborative grounding* and related concepts (Section 3). Section 4 plays with scenarios illustrating the pitfalls of being unable to ground collaboratively; we then turn to the language acquisition literature for insight into how humans do it (Section 5) and to the dialog systems literature to discuss what is known (Section 6). In Section 7, we reflect on the progress made and make recommendations, while in Section 8 we note possible objections to our account and conclude.

2 Motivation

Dialog and interactive systems is one of the most popular research areas in computational linguistics nowadays. But — unlike machine translation and information retrieval — deep learning approaches to it have had little impact on products that people use daily.¹ In 1991, cognitive scientist Brennan asked: *Why is it that natural language has yet to become a widely used modality of human/computer interaction?* (Brennan, 1991), and in 2020 the AI researchers de Vries, Bahdanau and Manning (de Vries et al., 2020) asked the same question yet again. In 1990, research on dialog systems used symbolic approaches; today neural generative models are favoured. Methods have changed, but the question remains the same.

Neural generative models offer flexibility, can be easily adapted to new domains, and require minimal domain engineering. But though they generate fluent responses (Serban et al., 2016), the result is often boring and repetitive (“I don’t know”) or they contradict themselves, or wander away from the

¹Almost all commercial dialog systems currently available seem to be based on pre-deep learning pipeline architectures (Takanobu et al., 2020) in spite of efforts such as the Alexa Prize (Ram et al., 2018; Gabriel et al., 2020).

topic of conversation (Li et al., 2016). De Vries et al (2020) note that neural generative models assume that the required data is available and appropriate. They say: *Ideally, the data used for training and evaluating should reflect the intents and linguistic phenomena found in real-world applications and be of reasonable size to accommodate modern data-intensive methods.* They remark that data quality and quantity are hard to reconcile, and that the community has prioritized quantity over quality; thus dialog systems are inadequate because datasets are poor. But ineffective use of dialog history seems to play a role too. Sankar et al (2019) study model sensitivity to artificially introduced context perturbations at test time. Working with multi-turn dialog datasets, they found that commonly used neural dialog architectures, like recurrent and transformer-based seq2seq models, are rarely sensitive to perturbations such as missing or reordered utterances, and word shuffling.

The ineffective use of dialog history often goes unnoticed because of the evaluation practices that are common nowadays. Automatic metrics such as BLUE, ROUGE, and so on, do not correlate well with human judgement, either for semantic preserving natural language generation or for dialog (Mathur et al., 2020). For human evaluation, it is not enough to show a few turns to the annotators (Liu et al., 2016). This does not measure how well the system is able to recover from its own mistakes; a human-in-the-loop evaluation that judges the overall interaction is needed for that (Walker et al., 1997).

Graphical user interfaces (GUIs), on the other hand, tend to get things right. We agree with Brennan (1998) that GUIs are more successful in everyday use than dialog systems (DSs) because GUIs enable collaborative grounding effectively while deep learning approaches to DSs do not.

3 What is collaborative grounding?

In this section we provide working definitions of collaborative grounding and other key concepts. To get the ball rolling, we take the *common ground* to be the commitments that the dialog partners have (explicitly or implicitly) agreed upon.

Collaborative grounding not symbol grounding

Collaborative grounding is the process of seeking and providing incremental evidence of mutual understanding through dialog; we view the ongoing exchange of speaker and hearer roles as fun-

damental to conversation (Benotti, 2010; Benotti and Blackburn, 2014). When the speaker believes that the dialog is on track, *positive evidence* of understanding is provided in different forms (depending on the communication channel) such as explicit acknowledgements and eye contact. *Negative evidence* of understanding signals that something needs to be negotiated before the dialog partners can commit — and negative evidence is ubiquitous:

Conversations with other people are rarely fluent and without mishap, and people do not expect them to be. (Brennan, 1991)

Mishaps lead to the need for *repair*. Repair is fundamental to *conversational analysis* (Schegloff et al., 1977; Schegloff, 2007), the linguistic study of language-driven social interactions. Together with the more psychologically oriented work of Clark and his colleagues (Clark and Wilkes-Gibbs, 1986; Clark, 1996), conversational analysis, is a key inspiration for the lines of research that we review in this paper.

We consider collaborative grounding to be distinct from symbol grounding (Harnad, 1990) though they interact in interesting ways (Larsson, 2018). *Symbol grounding* (or perceptual grounding, or language grounded in vision) is the set of capabilities that link symbols with perceptions; it is an important research area (Roy, 2005) in which accurate connections between systems’ linguistic representations and sensor data are often viewed as proof that a system means what it says. These connections are important for meaning, but we agree with De Vault et al. (2006) that *perceptual grounding is neither necessary nor sufficient to justify the attribution of linguistic meaning*. Human perception and memory are neither accurate nor stable, and different people have different abilities and limitations. Human meaning attributions do not rely on the accurate perceptions and perfect memory sought by symbol grounding, but on a *collaborative negotiation process* in which language speakers coordinate their perceptual memories and linguistic usage with other members of their communities. If a dialog system *commits* itself to negotiate its meanings collaboratively when perception and memory falter, then we claim that this gives grounds for assigning linguistic meaning to it. See Section 4 and A for examples.

Collaborative grounding: basic mechanisms

When people talk to each other, they tailor their

utterances to their partners. People can talk with friends, strangers, disembodied voices on the telephone, readers who will come along after they are gone, foreigners, children, and even dogs. Flexibility in tailoring utterances for a particular addressee has been documented even among the very young; five year olds have been observed to use more simple language and a different pitch range when talking to three year olds than they do talking to adults (Golinkoff, 1986). People adapt by initially estimating the common ground they share with a particular partner, by monitoring the positive and negative evidences of understanding (aka *grounding acts*) and by adapting their initial common ground estimate accordingly.

Alikhani and Stone (2020) explain that dialog systems can participate in collaborative grounding by ensuring they get attention and feedback from their users and tracking user state. Such *pragmatic mechanisms* have been explored, including those for dealing with problems related to joint attention (Koller et al., 2012; Koleva et al., 2015; Tan et al., 2020), engagement (Bohus and Horvitz, 2014; Foster et al., 2017), turn taking and incremental interpretation (Schlangen and Skantze, 2009; Selfridge et al., 2012; DeVault and Traum, 2013; Eshghi et al., 2015) corrections and clarifications (Villalba et al., 2017; Ginzburg and Fernández, 2010) and dialog management (DeVault and Stone, 2009; Selfridge et al., 2012). These mechanisms have been studied for different kinds of applications (Denis, 2010; Dzikovska et al., 2010, 2012). In Section 6 we discuss this research tradition; we believe it can provide top-down research guidance for research on dialog systems that commit to what they say.

Collaborative grounding: exposing limitations

As we said at the start, collaborative grounding is shaped by constraints that may not be explicit in successful dialog turns. Dialog partners use constraints to negotiate meaning, adding beliefs to the common ground through their joint attention to a real or imagined world. Further negotiation is required to unground a belief; unilateral belief withdrawal leads to a sense of commitment being brushed aside.

A dialog system is an unusual kind of linguistic partner. A user must figure out what its *limitations* are and then design her utterances with them in mind. Limitations may arise due to gaps in grammatical coverage, an unexpectedly narrow

domain vocabulary, or a lack of *common sense*; by common sense we here mean the basic knowledge and competencies required for successful navigation through a world full of objects, time, money, politeness, animals, people, and so on. One much-studied component of commonsense involves *causality* (Pearl, 2009) and the *frame problem* (Shanahan, 2016), the ability to figure out what changes and what does not change due to various kinds of (inter)action. Our point, however, is *not* that dialog systems need all these components of commonsense — it is rather that they need enough *interpersonal* common sense to not try and hide their inadequacies!

Often, however, these limitations are deliberately well hidden by the designers — who thereby unintentionally inhibit users from employing the powerful meaning-negotiation strategies available in a human-human dialog. Dialog systems may generate words they cannot understand, thus building false expectations of common ground in their dialog partner. Humans cannot adapt and ground collaboratively if a dialog system fails to *expose its own limitations*. In the next section we explore examples of such failures.

4 Test scenarios

This section uses playful scenarios to illustrate the implications of not learning to do collaborative grounding appropriately. We start with a text adventure scenario, then consider a multimodal scenario followed by a BlenderBot example, and conclude by contrasting these with the collaborative grounding typical of GUIs.

4.1 Dungeon intelligence

Consider the following interaction with the text game adventure AI Dungeon² powered by GPT-2 (Radford et al., 2019). The Webtext dataset used to train GPT-2 contains web content from Reddit links that have been rated as trustable by some Reddit users, probably including blogs, online news, and other discourse data.

In the example, when we detect a problem, we prompt Dungeon AI (DAI) with a negative grounding act (for example, *but before you only had seven, how many are there really?*). This makes the inconsistency with the previous turns explicit; in effect, we offer to negotiate. But the model blithely ig-

²<https://play.aidungeon.io/> accessed on 19/08/2020

nore the negative evidence and talks about something else:

DAI: You are wearing nondescript clothing, and you manage to escape without being recognized. You open the purse and find seven gold coins, which is good since you only planned to steal a little.

Human: How many coins are there in the purse?

DAI: You count the coins and find that you have eighty one coins, which is good since you planned to steal just a little. You think to yourself “I should be able to afford a room tonight.”

Human: But before you only had seven, how many are there really?

DAI: You head to the first inn you see, which seems to be a place where merchants stay. The desk clerk is a young woman who is probably only just thirteen or fourteen.

We posed this question several times but the model never tried to negotiate the inconsistency (for example, by replying: *there are 81 coins, I saw only some of them before*). In A there is a continuation of this dialog, illustrating other kinds of inconsistency and lack of collaborative grounding.

Bender and Koller (2020) argue that language models like GPT-2 are unable to ground language in the world: because they are only exposed to form, it is unreasonable to expect them to negotiate meaning. Others propose to tackle all kinds of grounding experimentally, by collecting datasets grounded in various modalities, developing models that learn from them, and using leaderboards (Linzen, 2020) and checklists (Ribeiro et al., 2020) to measure how effectively the model generalizations aligns with those of humans. So let’s turn to a multimodal task: visual dialog.

4.2 Visual dialog

Visual dialogs have long been a test-bed for natural language understanding. They played a prominent role in early work on natural language understanding (Winograd, 1972) and are now the focus of an active community investigating the interplay between computer vision and computational linguistics (Baldrige et al., 2018; Shekhar et al., 2019). Important progress has been obtained thanks to the recent release of datasets like VisDial (Das et al., 2017) and GuessWhat?! (de Vries et al., 2017); the former contains chit-chat conversations about an image whereas the latter is a visual game, hence its dialogs are goal oriented.

As we see in Figure 1, GuessWhat?! is a cooperative game: two players attempt to identify an object in an image. The *Questioner* has to guess the referent by asking yes/no questions; the *Oracle* knows the referent object and provides the answers. In the figure, the referent is the second woman from right to left of the group of four women in the back.

The Oracle model that includes the dialog history as an input (de Vries et al., 2017), correctly answers 9 out of 10 questions for the dialog in Figure 1; it only fails on question number 9, incorrectly answering “no”. But question 9 does not look particularly hard; why does it create a problem? *Because it trades on the common ground built by the previous dialog turns.* Humans use common ground here, as shallower methods fail (the referent is not salient in the image). Indeed — except for one turn — Figure 1 is just visual question answering; the system fails on precisely the question that *cannot* be correctly answered without using the common ground built by the previous turns. Is this error only due to the rarity of such dialog turns (as (Agarwal et al., 2020) argue), something which might be fixed by better datasets?

The answer may be “no”. Collaborative grounding in visual dialog has *two* characteristics that make it difficult for current machine learning methods: it is not only that turns requiring common ground may be infrequent in these datasets, in addition — as in our example — they may not be lexicalized. Common ground dependency is easier to spot when it is lexicalized with explicit pronouns (e.g. ‘them’ in “is it one of them?” in position 7). If the question “is it the second one?” had been asked in position 2, the model’s answer “no” would have been correct, as “second” would then refer to the second (highly salient) man in the row. But in position 9 the correct answer is “yes” due to the common ground built by turns 6 to 8. Common ground dependence may be less lexicalized in visual dialog as the visual modality tends to favor visual saliency to resolve ambiguities.

As we said in Section 3, common ground is the set of commitments that the dialog partners have agreed upon, while collaborative grounding is the process by which such commitments are made. Common ground dependence, as illustrated in turn 9 in Figure 1, may not be as frequent and as lexicalized in multimodal datasets as current data hungry models need (Agarwal et al., 2020; de Vries et al., 2020). But better datasets alone will not solve the



1. It is a person?	yes
2. It is a male?	no
3. Is it a female?	yes
4. Is she in the red dress?	no
5. She is in black?	yes
6. In the back from right to left you see four women?	yes
7. Is it one of them?	yes
8. Right to left, is it the first one, gray sweater?	no
9. Is it the second one?	yes
10. Does she look to have panty hose on?	yes

Figure 1: Human-human dialog on the Guesswhat?! referential task extracted from (de Vries et al., 2017). The example illustrates the linguistic variability present in this task.

problem: we need models that are trying to do collaborative grounding. That is, we need models that are not only led by the lexical cues in the dialog but also by pragmatic mechanisms reflected by timely exchanges between speaker and hearer. So let us turn to a model that has been exposed to such pragmatic mechanisms in its data.

4.3 BlenderBot

Facebook AI has built and open-sourced BlenderBot, which they call the largest-ever open-domain chatbot. It outperforms systems such as Google’s Meena (Adiwardana et al., 2020) in terms of engagement and also feels more human, according to human evaluators (Smith et al., 2020). BlenderBot attempts to combine different conversational skills including empathy, knowledge, and personality together in one system. The trained models are available for research (Smith et al., 2020) in different sizes and with different hardware requirements. The largest model has 9.4 Billion parameters, the middle sized version has 2.7 Billion.

Unlike the GPT-2 model we discussed earlier which is mostly trained on discourse data, BlenderBot is pre-trained with a large dataset of multi-party conversations extracted from the Reddit Pushshift dataset (Baumgartner et al., 2020). The dataset consists of free-form exchanges between multiple speakers and hearers where collaborative grounding is occurring. Thus the data on which BlenderBot models are pre-trained includes positive and negative evidence of understanding. BlenderBot models are then fine-tuned on dialogs that have been crowdsourced to exhibit empathy, knowledge about some particular topic, a consistent persona, and on a crowdsourced dataset that blends these abilities together (Smith et al., 2020).

The following interaction was generated using the middle sized model trained on Reddit Pushshift through July 2019 (Smith et al., 2020); we call the model BB2.7B.³

Human: Do you know Blender?

BB2.7B: Yes, I’ve used it a little bit. I am not very good at it, though.

Human: But I don’t mean the 3D software.

BB2.7B: Oh, I see. Yeah, I know a little bit about it. What do you like to do with it?

Human: I mean Blender, the chatbot. Have you used it?

*BB2.7B: Oh, yes, I have used it. I like to use it to make 3D models.*⁴

This fragment explores BlenderBot reaction to negative evidence (*but I don’t mean the 3D software*). It does not ignore it, as the Dungeon AI based on GPT-2 does. Indeed it replies with a coherent follow up and it includes a sentence that seems intended to acknowledge the misunderstanding (*Oh, I see*). However, the rest of the dialog shows that in spite of recognizing the structure of negative evidence, BlenderBot is unable to integrate negative grounding into the conversation consistently.

4.4 Graphical user interfaces

GUIs exploit graphical elements that mimic physical objects: we can point, drag and toss them in the trash bin. GUIs respond by updating immediately, thus the relationship between the user’s action and the graphical result is utterly clear. Even though GUIs are primarily graphical, they are also conversational and implement pragmatic mechanisms;

³The model is available at <https://parl.ai/projects/recipes/>

⁴See https://www.youtube.com/watch?v=-op_yOXBNEY for the full dialog.

indeed, their response is as timely and relevant as backchannels in human conversation (acknowledgments, nods, eye contact; see (Gravano and Hirschberg, 2011)). As they enable direct manipulation, they trivially solve the linguistic reference problem. They model common ground by tracking what is visible to the user. They model joint attention graphically through focus. And they do not suffer from the frame problem: consistency is carefully preserved in GUI design (Brennan, 1998); if you move something, it will stay there until somebody moves it back. They ask for both positive (e.g. ok) and negative (e.g. cancel) evidence for understanding. GUIs are good at exposing their own limitations and most users are good at adapting to them — some even overadapt and blame themselves for misunderstandings. For example, take the dialog box window in Figure 2. The system wants positive grounding evidence from the user but (confusingly) it does not offer the choice of giving negative grounding evidence (the conventional negative grounding label for buttons in GUIs is “cancel”).

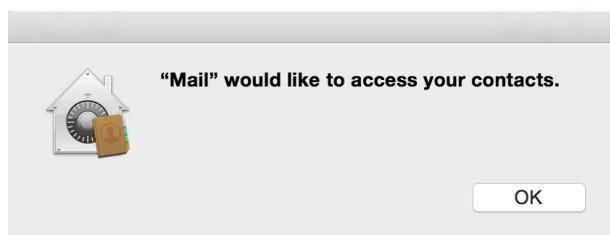


Figure 2: Sample dialog box not allowing negative evidence of grounding: there is no cancel button.

5 Human language acquisition

Clark (2001) presents evidence that collaborative grounding underpins the process of first language acquisition in babies. She argues that collaborative grounding offers a way of placing a new piece of the language at the center of *joint attention* of the language learner and her caregiver. In particular, it is through such pragmatic mechanisms that:

(a) *children solve the general mapping problem between form and meaning when offered new words; (b) they take up conventional terms for unfamiliar objects and events, and see how they relate to familiar terms; (c) they take up information about what they should have said when they have produced an erroneous utterance when offered reformulations.* (Clark, 2001)

Babies around one year old have been shown to perform *negative grounding acts* in order to *repair* a request that they made (e.g. “doll!”) whose *intention* (e.g. “getting the doll”) was *misunderstood* by an adult. Babies tend to do the repair act even when the request is satisfied by other means (maybe the frustrated adult gave the baby all the toys including the doll — yet the baby takes the doll and repeats “doll!”). In other words, they care that their intention is understood, not only satisfied (Ackermann et al., 2014; Tomasello et al., 2005). Tomasello et al argue that this is a basic ability, one required for the development of language and cognitive capabilities like belief attribution. Golinkoff describes this ability as follows:

Importantly, from their earliest forays into linguistic communication, infants engage in a “negotiation of meaning” in which they request clarification from the adult and produce communicative repairs for the adult when needed [...] Infants can and will persevere in the face of failure by altering their signals in creative non stereotypical ways. (Golinkoff, 1986)

Developmental psychologists have documented repeatedly that children with autism have difficulties signaling non-comprehension and making appropriate repairs to their own linguistic messages (Katherin et al., 1990). Deaf and hearing children have been found to employ different repair strategies. Deaf children were also more likely to revise utterances; hearing children more likely to provide cue repairs. When facing communication breakdown, both deaf and hearing children persisted effectively in clarifying utterances (Ciocci SR, 1998).

Allwood and colleagues have documented that grounding acts have a central role not only for first language acquisition but also for second languages (Allwood et al., 1991; Allwood, 1993, 1997; Allwood and Ahlsen, 1999).

6 Previous work, key insights

In this section we focus on previous work (and key insights) on collaborative grounding from research on human dialog analysis and dialog systems. We won’t cover work from robotics and symbol grounding; for that see e.g. (Roy and Reiter, 2005; Bohus and Rudnický, 2009; Bohus et al., 2012; Larsson, 2018).

Long before the deep learning era, dialog system researchers were aware that constructing common

ground *collaboratively* is a key task. A notable pioneer was Traum (1991; 2003) with his focus on dialog turns whose conversational role is to provide positive and negative evidence of grounding as tools for negotiating meaning. The interaction between dialog system and human computer interaction research was fruitful back then, but the tools available today for dealing with language variability (Mrksic et al., 2017) were not yet developed so systems were brittle (Allen et al., 2001).

DeVault and Stone (2009) showed that using predefined semantic tags (e.g. colors of objects) falls short for human dialog; people tend to invent new tags collaboratively as needed (e.g. using *greenish blue* to distinguish one object from a bluer one). This example could be described as “zooming into” the details of the available distinctions, but “zooming out” can occur when limitations are revealed. Saying *I am color blind* will (hopefully) shift the dialog away from reliance on color terms.

Rieser et al. (2010; 2011) and Georgila et al. (2005) propose using Wizard of Oz data instead of naturally occurring human-human dialog for training; this puts the spotlight on the necessary constraints and primes dialog systems with explicit grounding subdialogs for overcoming limitations, instead of restricting attention to the limitations of the channel through which the human crowd-sourcers interact. This enables data to be collected that makes explicit strategies for negotiating meaning, and could allow systems to learn particular collaborative grounding skills relevant for the dialog system task.

The surface form of explicit negotiations of meaning in dialog are frequently non-sentential utterances (Fernandez, 2006; Fernández et al., 2007). These include prototypical positive and negative evidence of grounding such as acknowledgements and clarification requests (Stoyanchev et al., 2013; Benotti and Blackburn, 2017), but also less-well-known forms such as self-corrections, rejections, modifiers and plain old questions and answers (Purver, 2004; Purver et al., 2018). Such work makes it evident that non-sentential utterances are not errors of performance and do not need “fixing” into sentential utterances. Ginzburg and Fernandez (2010) contributed detailed formalizations of how different evidences of grounding modified the public common ground and the private commitments of each dialog participant. A simple observation in (Ginzburg, 2012) does a big

job in illustrating how different written discourse and dialog can be: it is often said that the most common word in written discourse is ‘*the*’ while the most frequent word in naturally occurring conversations in the British National Corpus (Clear, 1993) is ‘*yes*’. This makes it evident that the contexts available to the dialog partners in the aftermath of an utterance are *not* identical. Positive acknowledgements (like ‘*yes*’) signal that the participants are synchronized and that collaborative grounding is proceeding smoothly (Denis et al., 2007).

In Section 3 we said that the ongoing exchange of speaker and hearer roles is fundamental to conversation. Schlangen and others have shown in detail how the ongoing exchange of these roles is so natural that we complete and correct each others sentences in *incremental* approaches to dialog (Schlangen and Skantze, 2009; DeVault et al., 2009; Baumann and Schlangen, 2012; DeVault and Traum, 2013; Kennington and Schlangen, 2017).

Hough and Schlangen (2017) argue that embodied dialog systems must ground the degree of uncertainty they have; that is they must *expose their limitations* as we argued in Section 3 and as GUI systems routinely do. They show that humans can reliably understand the level of uncertainty that a robot has and act accordingly. No complex natural language generation abilities are needed for this; negative grounding can be realized by perceivable hesitations in a physical act done by the robot.

Koller and colleagues explain how the joint attention of the conversational participants can be considered and manipulated in order to correct the common ground. In particular, they show how listener attention can be manipulated when there is co-presence through speech and gaze (Koller et al., 2012; Koleva et al., 2015) and through emphasis in the text (“press the RED button”) when co-presence is not possible (Villalba et al., 2017).

In sum, *the possibility of making mistakes and collaboratively recovering from them is one of the key pragmatic mechanisms for grounding meaning*. Much is already understood about this process, and in Section 7 we recommend using this work to motivate top down advances in dialog and interactive systems that would complement the bottom up approaches described in Section 2.

7 Moving forward

In *What Computers Can’t Do* (1972), Dreyfus drew on the ideas of philosophers like Merleau-

Ponty (Merleau-Ponty, 1962) and Heidegger (Heidegger, 1993) to criticize symbolic AI. Dreyfus emphasized the embodied capability of *knowing how*, rather than the abstract propositional *knowing that* typical of symbolic AI; he did not anticipate that AI would find plausible methods (deep neural nets, embodied robotics, distributional semantics) for exploring knowing how.

Dreyfus' criticisms may seem obvious in retrospect, but it is useful to recall another philosopher that he cited. In *Philosophical Investigations*, Ludwig Wittgenstein (1953) critiqued earlier approaches to language and meaning (including his own) for failing to take the collaborative aspect of language into account.

Wittgenstein's later work foregrounds the importance of *social interaction*. In Sections 3 and 4 we remarked that collaborative grounding is more than symbolic/perceptual grounding, and claimed that the crucial missing component is provided by social interaction. In Section 5 we saw that human children are born into is a complex world of agents, relationships, affect and much else beside. Moreover (as the child soon learns) it is a world in which interesting others collaborate with the help of a malleable system called language. This system is capable of expressing multiple types of meaning — symbolic, scientific, social — which Wittgenstein summed up anthropologically: language was a form of life (*Lebensform*).

Wittgenstein's ideas are inspiring, but we don't need to look so far back: the work we reviewed in Sections 5 and 6 shows that social aspects of language use are central to the roots of meaning, and cannot be an afterthought in dialog models. This leads to our recommendations.

Appreciate socially grounded dialog datasets

Dialog datasets are scarce as most naturally created dialog corpora cannot be shared due to privacy issues. Matters are different in written discourse, which has made great progress thanks to vast corpora from news services, Wikipedia and the like. We agree with de Vries et al (2020) that more effort should be put into designing what they call *ecologically valid* datasets for dialog. We agree that datasets should avoid four common issues: synthetic language, artificial tasks, not working with prospective users, and single-turn interfaces. We add two items to this wish-list: (1) dialog grounded in at least one modality and (2) dialog grounded in a (symmetric or asymmetric) social collaboration.

Regarding (1), there is much work that has been done on data collection in the visual (i.e. seeing) modality (Baldridge et al., 2018) and more recently in the kinaesthetic (i.e. moving) modality (Weihs et al., 2020). The work we reviewed in Section 6 is a good starting point for designing tasks that also consider (2) for various modalities: auditory (hearing) (Schlangen and Skantze, 2009), tactile (touching) (Hough and Schlangen, 2017), kinaesthetic (moving) (Foster et al., 2014), and visual (seeing) (Koleva et al., 2015). Dialogs grounded in social collaboration can be symmetric or asymmetric. In symmetric dialogs, the dialog roles are socially exchangeable (e.g. (Illykh et al., 2019; Haber et al., 2019)). In asymmetric dialogs the roles of the dialog partners are different due to expertise, social power or for some other reason. Typical examples include instruction dialog (Anderson et al., 1991) and medical or technical support (Janarthanam and Lemon, 2009).

Our final recommendation regarding datasets is to let humans collaborate freely during the first round of data collection but then to collect more restricted datasets on the same task once the limitations of the model are clear. That is: let people adapt to its limitations and collect further data that reflects the collaborative adaptation mechanisms.

Use datasets better Better data alone may not be enough; we need to consider models that explicitly tackle collaborative grounding. Examples of such work that builds on fundamental research (like that reviewed in Section 6) already exists. Consider, for example, the following exchange extracted from a human-human dataset (Andreas et al., 2020) collected through a Wizard of Oz methodology:

User: What time is my planning meeting?

Agent: You meet with Grace at noon.

User: Sorry, I meant all-hands.

Agent: Your all-hands meeting is at 2:30 pm.

The fragment contains the negative grounding act *Sorry, I meant all-hands*; which requires access to the user's previous utterance to obtain the revised intention: *What time on Tuesday is my all-hands meeting?*. Promising results have been obtained using hybrid learned-symbolic dialog systems that explicitly model the intention by grounding it into an application domain. For example (Andreas et al., 2020) represent intentions (including grounding acts) as programs that modify the common ground; (El Asri et al., 2017) track the com-

mon ground using frames, (Lison and Kennington, 2016) do so using Bayesian networks, and (Ultes et al., 2018) using entities. Such approaches have been rather marginalized in favor of more shallow ones.

Interact with models in order to test them

Asking someone “do you need history to answer this question?” is not the same as answering it correctly without history. Most human evaluation of dialog systems are about self perception, not about actually performing an action. As is done in Human Computer Interaction, let people interact with models and then rate them; do not just offer pairs of turns. Dialog is *not* the concatenation of pairs of dialog turns, as noted by (Walker et al., 1997; Schlangen and Skantze, 2009; Ginzburg, 2012; Agarwal et al., 2020) and others (see Section 6).

Focus on error recovery, not error avoidance. Explore your dataset thinking about the *constraints* present in the dialog system you are building: Can they be learned from the data that you have? What are the limitations that you know your system will have? How can you expose these limitations to the dialog partner so that she can adapt to them? Will a Wizard of Oz setup in which a human simulates the system limitations help here?

The metric for evaluating a dialog system should not (only) be accuracy on some static dataset, but also: *how many mistakes you cannot recover from when interacting with a potentially adversarial human being*. Other areas of NLP are already using such evaluations (Nie et al., 2020).

Design with collaborative grounding in mind

Many deep learning dialog systems differ from simple question answering approaches in recording the dialog history up to some limit, usually dictated by the number of tokens that can be reasonably encoded as model input (e.g. (Agarwal et al., 2020; Smith et al., 2020)). Which leads to a question: if collaborative grounding occurs in the conversations on which these models are trained, what exactly is missing in (say) BlenderBot, which was trained on Reddit conversations and has a dialog memory of at least a few turns? Is it that current training approaches do not capture this skill? For example could it be that the pre-trained model could exhibit such a skill but that fine-tuning to given tasks (on datasets where collaborative grounding does not occur) wipes it out? Or is it that something more is needed? We believe that attention needs to be

paid to the collaborative grounding mechanisms reviewed in Section 6.

Don’t leave the social aspects till last Bisk et al (2020) give a detailed description of milestones that must be passed to achieve what they call *truly contextual language understanding*. They argue that we are currently close to the written word milestone, the next milestone being perception, the following one embodiment, with the final one being the social. This may be motivated by the fact that humans evolved through perception and embodiment before developing language.

However language is not just about words and their semantics; it is primarily about people, their intentions and their interactions. We urge the community *not* to leave the social aspects of language for the future but to (re)consider existing work on collaborative grounding (Sections 3 and 6) and language acquisition (Section 5). We need to face up to language as *lebensform* right from the start.

8 Conclusion

Possible objections range from the practical (*new datasets are expensive; old datasets are too small for deep learning*) through the theoretical (*much coherent dialog does not require conversational history, task-oriented dialog may requires collaborative grounding but chit chat does not*) to the technical (*grounding acts cannot be identified automatically; deep learning models can learn to do some collaborative grounding without explicitly representing grounding intentions*). But lack of space precludes further discussion, so with these noted, we conclude by restating our basic point.

The idea that simply training models on large amounts of successful dialog as is currently done will lead to human-analogous collaborative grounding seems dubious. As we have described and illustrated, human grounding has to do not only with successful exchanges and positive grounding, but also with negative grounding. Negative grounding includes pragmatic mechanisms by which we negotiate what we actually mean. It is about errors and corrections, about exposing limitations. Until we figure out how to train these, our dialog models will lack a crucial capability: commitment to their own meanings.

Acknowledgments

We thank the anonymous reviewers and area chair for their detailed reviews and insightful comments.

References

- Hermann Ackermann, Steffen R Hage, and Wolfram Ziegler. 2014. Brain mechanisms of acoustic communication in humans and nonhuman primates: An evolutionary perspective. *Behavioral and Brain Sciences*, 37:529–546.
- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. [Towards a human-like open-domain chatbot](#). *arXiv*, abs/2001.09977.
- Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. ACL.
- Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim Marks, Chiori Hori, Peter Anderson, Stephan Lee, and Debi Parikh. 2019. Audio visual scene-aware dialog. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7550–7559.
- Malihe Alikhani and Matthew Stone. 2020. Achieving common ground in multi-modal dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 10–15. ACL.
- James F. Allen, Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2001. Toward conversational human-computer interaction. *AI Magazine*, 22(4):27–37.
- Jens Allwood. 1993. Feedback in second language acquisition. In *Adult Language Acquisition: Cross Linguistic Perspectives, II: The Results*, pages 196–235. Cambridge University Press.
- Jens Allwood. 1997. Notes on dialog and cooperation. In *IJCAI-97 Workshop on Collaboration, Cooperation and Conflict in Dialogue Systems*, pages 8–10.
- Jens Allwood and Elisabeth Ahlsen. 1999. Learning how to manage communication with special reference to the acquisition of linguistic feedback. *Journal of Pragmatics*, 31:1353–1389.
- Jens Allwood, Joakim Nivre, and Elisabeth Ahlsen. 1991. On the semantics and pragmatics of linguistic feedback. Technical Report 64, Gothenberg Papers in Theoretical Linguistics, Sweden.
- Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and Speech*, 34:351–366.
- Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitriy Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.
- Jason Baldridge, Tania Bedrax-Weiss, Daphne Luong, Sridhar Narayanan, Bo Pang, Fernando Pereira, Radu Soricut, Michael Tseng, and Yuan Zhang. 2018. Points, paths, and playscapes: Large-scale spatial language understanding tasks set in the real world. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 46–52. ACL.
- Timo Baumann and David Schlangen. 2012. IN-PRO.iSS: A component for just-in-time incremental speech synthesis. In *Proceedings of the ACL 2012 System Demonstrations*, pages 103–108, Jeju Island, Korea. ACL.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 830–839. AAAI Press.
- Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. ACL.
- Luciana Benotti. 2010. *Implicature as an interactive Process*. Ph.D. thesis, Université de Lorraine. Supervised by Patrick Blackburn.
- Luciana Benotti and Patrick Blackburn. 2014. Context and implicature. In Patrick Brézillon and Avelino J. Gonzalez, editors, *Context in Computing - A Cross-Disciplinary Approach for Modeling the Real World*, pages 419–436. Springer.
- Luciana Benotti and Patrick Blackburn. 2017. Modeling the clarification potential of instructions: Predicting clarification requests and other reactions. *Computer Speech Language*, 45:536–551.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020.

- Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8718–8735. ACL.
- Dan Bohus and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction*, page 2–9. ACM.
- Dan Bohus, Ece Kamar, and Eric Horvitz. 2012. Towards situated collaboration. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pages 13–14. ACL.
- Dan Bohus and Alexander I. Rudnicky. 2009. The ravenclaw dialog management framework: Architecture and systems. *Computer Speech and Language*, 23(3):332–361.
- Susan E Brennan. 1991. Conversation with and through computers. *User modeling and user-adapted interaction*, 1(1):67–86.
- Susan E. Brennan. 1998. The grounding problem in conversations with and through computers. In S. R. Fussell and R. J. Kreuz, editors, *Social and Cognitive Approaches to Interpersonal Communication*, pages 201–225. Lawrence Erlbaum, Hillsdale, NJ.
- Hyundong Cho and Jonathan May. 2020. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2398–2413. ACL.
- Baran J Ciocci SR. 1998. The use of conversational repair strategies by children who are deaf. *American Annals of the Deaf*, 143(3):235–45.
- Eve Clark. 2001. Grounding and attention in the acquisition of language. In *Proceedings of the 37th meeting of the Chicago Linguistic Society*, pages 95–116.
- Herbert Clark. 1996. *Using Language*. Cambridge University Press.
- Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1 – 39.
- Jeremy H. Clear. 1993. *The British National Corpus*, pages 163–187. MIT Press, Cambridge, MA, USA.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.
- Alexandre Denis. 2010. Generating referring expressions with reference domain theory. In *Proceedings of the Sixth International Natural Language Generation Conference*. ACL.
- Alexandre Denis, Guillaume Pitel, Matthieu Quignard, and Patrick Blackburn. 2007. Incorporating asymmetric and asynchronous evidence of understanding in a grounding model. In *11th Workshop on the Semantics and Pragmatics of Dialogue-DECALOG 2007*.
- David DeVault, Iris Oved, and Matthew Stone. 2006. Societal grounding is essential to meaningful language use. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, pages 747–754. AAAI Press.
- David DeVault, Kenji Sagae, and David Traum. 2009. Can I finish? learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue*. ACL.
- David DeVault and Matthew Stone. 2009. Learning to interpret utterances using dialogue history. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 184–192. ACL.
- David DeVault and David Traum. 2013. A method for the approximation of incremental understanding of explicit utterance meaning using predictive models in finite domains. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1092–1099. ACL.
- Hubert Dreyfus. 1972. *What Computers Can’t Do*. MIT Press.
- Myroslava O. Dzikovska, Peter Bell, Amy Isard, and Johanna D. Moore. 2012. Evaluating language understanding accuracy with respect to objective outcomes in a dialogue system. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 471–481. ACL.
- Myroslava O. Dzikovska, Johanna D. Moore, Natalie Steinhäuser, and Gwendolyn Campbell. 2010. The impact of interpretation problems on tutorial dialogue. In *Proceedings of the 2010 Conference of the Association for Computational Linguistics*, pages 43–48. ACL.
- Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. ACL.
- Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. ACL.

- Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matthew Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 261–271, London, UK. ACL.
- Raquel Fernandez. 2006. *Non-sentential utterances in dialogue: Classification, resolution and use*. Ph.D. thesis, University of London.
- Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.
- Mary Ellen Foster, Andre Gaschler, and Manuel Giuliani. 2017. Automatically classifying user engagement for dynamic multi-party human-robot interaction. *International Journal of Social Robotics*, 9(5):659–674.
- Mary Ellen Foster, Simon Keizer, and Oliver Lemon. 2014. Towards action selection under uncertainty for a socially aware robot bartender. In *ACM/IEEE international conference on human-robot interaction*, pages 158–159.
- Raefer Gabriel, Yang Liu, Anna Gottardi, Mihail Eric, Anju Khatri, Anjali Chadha, Qinlang Chen, Behnam Hedayatnia, Pankaj Rajan, Ali Binici, Shui Hu, Karthik Gopalakrishnan, Seokhwan Kim, Lauren Stubel, Kate Bland, Arindam Mandal, and Dilek Hakkani-Tür. 2020. Further advances in open domain dialog systems in the third alexa prize socialbot grand challenge. *Proceedings of the Alexa Prize*.
- Kallirroi Georgila, James Henderson, and Oliver Lemon. 2005. Learning user simulations for information state update dialogue systems. In *Ninth European Conference on Speech Communication and Technology*.
- Jonathan Ginzburg. 2012. *The Interactive Stance*. Oxford Press.
- Jonathan Ginzburg and Raquel Fernández. 2010. Computational models of dialogue. In Alexander Clark, Chris Fox, and Shalom Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*. Blackwell.
- Roberta Golinkoff. 1986. ‘I beg your pardon?’: the preverbal negotiation of failed messages. *Journal of child language*, 13:455–76.
- Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25(3):601–634.
- Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910. ACL.
- Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335 – 346.
- He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1766–1776, Vancouver, Canada. ACL.
- Martin Heidegger. 1993. The end of philosophy and the task of thinking. *Basic Writings*, pages 427–449.
- Julian Hough and David Schlangen. 2017. It’s not what you do, it’s how you do it: Grounding uncertainty for a simple robot. In *Proceedings of the International Conference on Human-Robot Interaction*, pages 274–282. ACM.
- Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meet up! a corpus of joint activity dialogues in a visual environment. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue*, London, United Kingdom.
- Srinivasan Janarthnam and Oliver Lemon. 2009. Learning lexical alignment policies for generating referring expressions for spoken dialogue systems. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 74–81. ACL.
- Loveland Katherin, Robin McEvoy, and Belgin Tunali. 1990. Narrative story telling in autism and down’s syndrome. *British Journal of Developmental Psychology*, 8(1):9–23.
- Casey Kennington and David Schlangen. 2017. A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech and Language*, 41:43–67.
- Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 6495–6513. ACL.
- Nikolina Koleva, Martín Villalba, Maria Staudte, and Alexander Koller. 2015. The impact of listener gaze on predicting reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 812–817. ACL.
- Alexander Koller, Konstantina Garoufi, Maria Staudte, and Matthew Crocker. 2012. Enhancing referential success by tracking hearer gaze. In *Proceedings of the 13th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 30–39. ACL.

- Dimosthenis Kontogiorgos, Elena Sibirtseva, and Joakim Gustafson. 2020. Chinese whispers: A multimodal dataset for embodied language grounding. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 743–749, Marseille, France. European Language Resources Association.
- Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–595, Minneapolis, Minnesota. ACL.
- Staffan Larsson. 2018. Grounding as a side-effect of grounding. *Topics in Cognitive Science*, 10(2):389–408.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 110–119. ACL.
- Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Meeting of the Association for Computational Linguistics*, pages 5210–5217. ACL.
- Pierre Lison and Casey Kennington. 2016. OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *Proceedings of ACL-2016 System Demonstrations*, pages 67–72, Berlin, Germany. ACL.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132. ACL.
- Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997. ACL.
- Maurice Merleau-Ponty. 1962. *Phenomenology of Perception*. Routledge.
- Nikola Mrksic, Ivan Vulic, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gasic, Anna Korhonen, and Steve J. Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.
- Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. ACL.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901. Association for Computational Linguistics.
- Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, USA.
- Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King’s College, University of London.
- Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in Cognitive Science*, 10(2):425–451.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, Open AI.
- Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrew. 2018. [Conversational AI: the science behind the alexa prize](#). *arXiv*, abs/1801.03604.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912. ACL.
- Verena Rieser and Oliver Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer Science & Business Media.
- Verena Rieser, Oliver Lemon, and Xingkun Liu. 2010. Optimising information presentation for spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1009–1018. ACL.
- Deb Roy. 2005. Grounding words in perception and action: Computational insights. *Trends in cognitive sciences*, 9:389–96.
- Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1):1 – 12.

- Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37. ACL.
- Emanuel Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*, volume 1. Cambridge University Press.
- Emanuel Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.
- David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 710–718. ACL.
- Ethan O. Selfridge, Iker Arizmendi, Peter A. Heeman, and Jason D. Williams. 2012. Integrating incremental speech recognition and POMDP-based dialogue systems. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–279, Seoul, South Korea. ACL.
- Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth Conference on Artificial Intelligence*, page 3776–3783. AAAI Press.
- Murray Shanahan. 2016. The Frame Problem. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, spring 2016 edition. Metaphysics Research Lab, Stanford University.
- Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and Guess-What. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2578–2587. ACL.
- Todd Shore, Theofronia Androulakaki, and Gabriel Skantze. 2018. KTH tangrams: A dataset for research on alignment and conceptual pacts in task-oriented dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents’ ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030. ACL.
- Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. 2013. Modelling human clarification strategies. In *Proceedings of the SIGDIAL 2013 Conference*, pages 137–141, Metz, France. ACL.
- Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. 2020. Is your goal-oriented dialog model performing really well? empirical analysis of system-wise evaluation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–310. ACL.
- Xiang Zhi Tan, Sean Andrist, Dan Bohus, and Eric Horvitz. 2020. Now, over here: Leveraging extended attentional capabilities in human-robot interaction. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, page 468–470, New York, NY, USA. ACM.
- Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691.
- David Traum. 1991. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester.
- David Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In Jan van Kuppevelt and Ronnie Smith, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer.
- Stefan Ultes, Paweł Budzianowski, Iñigo Casanueva, Lina M. Rojas-Barahona, Bo-Hsiang Tseng, Yen-Chen Wu, Steve Young, and Milica Gašić. 2018. Addressing objects and their relations: The conversational entity dialogue model. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 273–283. ACL.
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 673–683. ACL.
- Martín Villalba, Christoph Teichmann, and Alexander Koller. 2017. Generating contrastive referring expressions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Vancouver, Canada. ACL.
- Harm de Vries, Dzmitry Bahdanau, and Christopher D. Manning. 2020. [Towards ecologically valid research on language user interfaces](#). *arXiv*, abs/2007.14435.
- Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. [Talk the walk: Navigating new york city through grounded dialogue](#). *CoRR*, abs/1807.03367.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4475. IEEE Computer Society.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.

Luca Weihs, Jordi Salvador, Klemen Kotar, Unnat Jain, Kuo-Hao Zeng, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2020. *Allenact: A framework for embodied ai research*. *arXiv*.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3:1–191.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell.

A More examples from the test scenarios

In this section we present more examples that illustrate the basic points made in the test scenarios presented in Section 4 of the paper.

A.1 Dungeon Intelligence

This is a continuation of the dialog with the Dungeon AI (DAI) game (mode fantasy, character wizard) presented in Section 4. The example in the main paper illustrated numeric inconsistency. In this extended example we illustrate also gender inconsistency (the clerk is first a girl and then a man) and then causal inconsistency (after the food is brought the clerk says they do not serve food, but finally food is available).

DAI: You are wearing nondescript clothing, and you manage to escape without being recognized. You open the purse and find seven gold coins, which is good since you only planned to steal a little.

Me: How many coins are there in the purse?

DAI: You count the coins and find that you have eighty one coins, which is good since you planned to steal just a little. You think to yourself “I should be able to afford a room tonight.”

Me: but before you only had seven, how many are there really?

DAI: You head to the first inn you see, which seems to be a place where merchants stay. The desk clerk is a young woman who is probably only just thirteen or fourteen.

Me: I count the coins again

DAI: You count the coins and there are exactly 32 coins there. Hmm, makes it easy to know how much to give to beggars.

Me: I am hungry, I buy a burger.

DAI: You walk over to the counter and order a Whopper with cheese.

Me: I look at your burger.

DAI: It looks like meat. “A burger of goat meat and a mug of ale please.” The man nods and quickly brings you the food. “Thanks.”

Me: I am surprised, I thought the clerk was a girl.

DAI: “Sorry kid, I don’t serve food.”

Me: I am confused, I ask “Do you serve food or not?”

DAI: “Sure, if you have the coin.” He gives you the food. The food is delicious. You haven’t eaten anything so good in a while.

As mentioned in Section 4, it is unsurprising that language models trained only on form are unable to perform collaborative grounding. So in the next subsection we further explore a widespread dataset (de Vries et al., 2017) to look for evidence of collaborative grounding in a multimodal task.

A.2 Visual dialog

Figure 3a, 3b and 3c are random dialogs extracted for the Guesswhat?! dataset (de Vries et al., 2017). In all dialogs, the meaning of the question is correctly interpretable without the previous dialog. In Figure 3a it seems that question 7 *On the left side?* is dependent of the previous turn, and means *On the left side of the boy in the backwards baseball cap?*. However it can be answered correctly with an absolute interpretation without context such as *On the left side of the picture?*. Similarly, in Figure 3b it seems that question 7 *Is it touching the right edge?* should be interpreted as *Is the carrot touching the right edge?*. However, it can correctly answered in a context independent fashion interpreting it as *Is it something touching the right edge?*. Finally, in Figure 3c even the elliptical question 6 *Partially visible?* can be answered correctly without considering the previous turns of the dialog.

The ellipsis can be resolved by changing the question to *It is the one partially visible?*. Even though the “partially visible” criterion hold of more than one potential referent, it is only necessary to know it is true for the target to answer it correctly.

From the Oracle perspective, the whole interaction can be solved without access to the dialog



Question	Answer
1. Is it alive?	yes
2. Is it the baseball player?	no
3. Is it a fan?	yes
4. Is he standing?	no
5. In he wearing a backwards baseball cap?	no
6. Next to the boy in the backwards baseball cap?	yes
7. On the left side?	yes

(a) Question 7 can be interpreted with respect to question 6 as “on the left side of the boy with the backwards baseball cup”.



Question	Answer
1. Is it the plate?	no
2. Is it food?	yes
3. Is it a little piece of food?	yes
4. Is it carrot?	yes
5. Is it on the left side?	no
6. Is it on the far right?	yes
7. Is it touching the right edge?	yes

(b) Question 7 could be enriched to “is it the little piece of carrot touching the right edge”, considering to questions 3 and 4.



Question	Answer
1. Is it a person?	yes
2. Are they wearing blue?	yes
3. Are they one of the three in the front?	yes
4. Is it a man?	no
5. A woman wearing shorts?	yes
6. Partially visible?	yes

(c) Question 6 could be enriched by the previous dialog as “is it a partially visible woman in blue wearing shorts”.

Figure 3: Sample image and dialog from the GuessWhat?! dataset (de Vries et al., 2017). The green mask marks the dialog referent. Questions seem dialog dependent but they can be answered correctly in isolation.

history. There is no need for the Oracle to engage in collaborative grounding for these dialogs.

A.3 Graphical User Interfaces

In this section we illustrate another parallelism between a bad graphical user interface (GUI) and a bad dialog system. A dialog systems that does not expose its limitations and acts as if it was able to detect more intentions that it can actually do, may be compared to a GUI that offers too many options as illustrated in Figure 4.

Both of these interfaces, the graphical and the conversational, overwhelm the user with options, many of which may turn out to be useless. Moreover, a GUI that uses icons that are unfamiliar to

the target user, or are hard to recognize, suffers from the same problems as a conversational system trained with a different language variety or expertise than that of the user. This last problem can be overcome through collaborative grounding. In particular, if users need to use this GUI, they can ground the meaning of the strange icons through interaction — all the needed information will be exposed by the GUI. We illustrate both points in Figure 4, a GUI cluttered with strange icons.

B Examples of collaborative grounding

Building datasets is expensive. In this section we briefly describe dialog datasets already built that include collaborative grounding phenomena. An

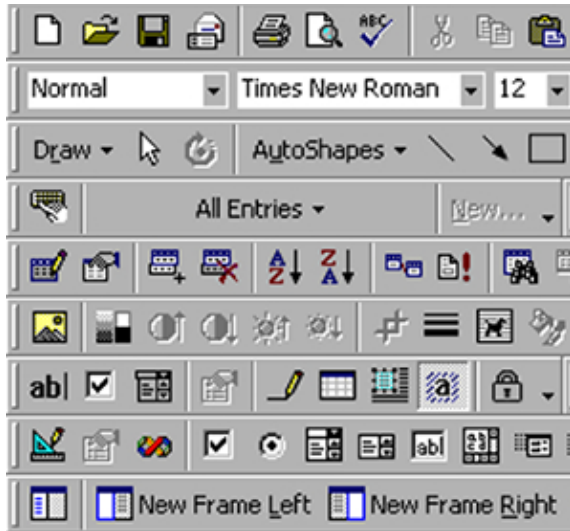


Figure 4: A cluttered GUI that contains icons which are not meaningful for the user

interesting direction of future research would be to analyze and quantify the collaborative grounding acts in these 2 datasets in the spirit of (Fernandez, 2006; Fernández et al., 2007) and other work reviewed in Section 6 of the paper.

The following is an example from (Clark and Wilkes-Gibbs, 1986) with a trial noun phrase collaboratively grounded between S and J. This dataset motivated the creation of larger datasets on the same task (Shore et al., 2018).

S: *The small blue cap we talked about before?*
 J: *The one with yellow dots?*
 S: *Yeah*

The following example from the Photobook dataset (Haber et al., 2019) starts with a trial noun phrase from A who referred to a TV as a computer, which is collaboratively grounded between A and B. The authors propose to integrate grounding acts to a model based on reference chains.

A: *Man with dog on lap looking at his computer?*
 B: *I don't have that, but could it be a TV in yours?*
 Mine has a man sitting with his dog watching TV.
 A: *yes, TV - sorry!*
 B: *Okay.*

This example shows two dialog participants collaboratively grounding their position in a map in the MeetUp corpus (Ilinykh et al., 2019). The map contains pictures of the different rooms the participants can be in. They coordinate the position by describing the rooms. The task is designed to be symmetric, so both participants can contribute equally.

Other datasets with a symmetric task are Mutual Friends (He et al., 2017) and Light (Urbanek et al., 2019) and others (Cho and May, 2020).

B: *I think I am in a basement*
 B: *i think i might be too*
 A: *maybe not though*
 A: *wood panel?*
 A: *two doors?*
 B: *there's a tan couch, and a tan loveseat/chair*
brown coffee table. bar. tv
 B: *nope, different room*
 A: *ok i am not there*

The following dialog fragment was collected between crowdsourcers acting as a tourist and a guide (de Vries et al., 2018). By looking at a map the guide had to accomplish the goal of guiding the tourist to a given location in a city. The tourist had access to a navigation street view and provided feedback to the guide about what he did and saw. The task is designed to be asymmetric so collaboration is limited to the roles each participant plays. The guide can make mistakes because the map information is incomplete. The tourist takes a more active role when the guide makes a mistake as illustrated below.

Guide: *Ok, turn left then go straight up that road*
 Guide: *There should be shops on two of the corners but you need to go to the corner without a shop.*
 Tourist: *on my left is Radio city Music hall*
 Tourist: *I can't go straight any further.*
 Guide: *ok. turn so that the theater is on your right.*
 Guide: *then go straight*
 Tourist: *That would be going back the way I came*
 Guide: *yeah. I was looking at the wrong bank*

Other datasets collected for asymmetric tasks are (Eric et al., 2017; Kottur et al., 2019; Alamri et al., 2019; Kim et al., 2019; Narayan-Chen et al., 2019; Kontogiorgos et al., 2020). All these datasets show collaborative grounding phenomena and are promising contributions for the development of models that can learn to collaboratively ground meaning.